Food survey using exploratory data analysis.

RAMYASRI, R., ISHASANJIDA, S., PARASA, D. and BANO, S.

2019

Food Survey using Exploratory Data Analysis

Rayapati RamyaSri
Dept of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
rayapatiramyasri199@gmail.com

Shaik IshaSanjida
Dept of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram,India
ishashaik77@gmail.com

Dhanush Parasa
Dept of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
cody20734@gmail.com

Shahana Bano
Assoc Professor
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
shahanabano@icloud.com

***Abstract*—We are well aware of the many problems that our current generations are facing. From all these new enhancements in the real world it has been quite hard for them to keep up with everything evolving around them. Keeping all this in mind they work day in and out to make sure that their knowledge on their surroundings up to date, however we believe that they fail to properly take care of themselves in the process. No matter how much a certain individual may withstand in terms of workload, stress, or other mental & emotional barriers our physical body will always be the key aspect to overcoming them. Most people believe that working out and maintaining physical fitness are the major aspects to sustain a healthy physical form but they simply overlook the most important aspect which are their eating habits. Although our body may be physically fit, the nourishment of our body depends on the eating styles that we follow on a day to day basis. Food is what nourishes our body with most of the proteins & minerals that we require, without it we wouldn't be able to accomplish much. On conducting a worldwide research on people's lifestyles we were able to conclude that over the past 33 years the obesity rate among human beings has increased by a mere 27.5%. What seems to be the most thoughtful yet intriguing fact is that although many people are overweight as well as obese they still believe that their eating habits are healthy. Most people are living in the dilemma of the fact that they maintain a healthy lifestyle. We aim to study the views on a healthy lifestyle as per the norms of our current generation. We would like to analyse their daily eating habits as well as their own thoughts on their lifestyle. So the question that remains is…**

"What exactly is a Healthy Eating Lifestyle?"

***IndexTerms*: DataExplorer,CorrelationAnalysis,Principle Component Analysis**

I. INTRODUCTION

Our project mainly focuses on analysing the lifestyles in terms of food on our current generation. In order for us to accomplish this we have conducted a large survey which consisted mostly upon students from our college as well as a few of the college in our locality. In order to make our process much simpler we were able to use **R-programming** for the analysis of all of our Data…

' **R**' is an open world software which was developed in 1993 for statistical computing and graphics. It is one of the most used software's by data analysts as well as data miners so that they can develop statistical software's.' **R** ' is a language that comes with many inbuilt libraries that can be used in various scenarios. However this language is well known for the following algorithms; Linear regression, Logistic regression, Decision tree, SVM, Naïve Bayes, KNN, K-means, and Random Forest. These may only be a few of its main algorithms but they are most used by many of its users. We have implemented ' **R** ' within our project to analyze all of our data sets from our survey. It was also able to clearly plot all of our data using its in-built graphical libraries.

II. PROCEDURE

Our project focused on 3 major stages. The first stage consisted of us gathering the required Data for our Study. The next two consist of us importing the collected data into our program and having it analysed. Once we have successfully analyzed our data we should be able to implement the various libraries within ' **R '** so that we can also have statistical graphs drawn for a visual representation. Let's take a closer look at each step in our process…

A) Collection of Data:

In order for us to commence our study on the lifestyles of the current generation we needed data sets to work with. We needed data sets that would help us understand 2 major aspects…

i) The Daily Eating Habits of a person.

ii) The person's outlook on whether their eating style is healthy or not.

Keeping these two key aspects in mind we created an online survey using google forms for students to fill out. The survey mainly focused on a person's daily eating habits focusing on what types of food they consumed. From these habits we ended our survey with a simple question on their views or opinions on whether their lifestyles are healthy.Once our survey was completed we were able to output our data in the form of an Excel sheet. The sheet had all the questions placed along the x-

axis while each entry was considered on the y-axis. Once we have properly saved our excel sheet on our computer we can now progress to the next stage of implementation and Analysis.

The attributes that we gathered from our survey form are Age, Gender, Diet Plan, time you wake in week days, time you wake in week-end, prefer as soon as u wake, Breakfast in a Rush, skip Breakfast, more in Breakfast, Breakfast in Restaurant, Breakfast in Canteen, Breakfast in Street, After your Breakfast, break time, Lunch Time, food in Lunch, type of food in Lunch, skip Lunch, Snacks you eat more, often you have your Snack, Dinner Time, eat more in Dinner, skip your Dinner, after Dinner, kind of food you prefer more and other values for attributes.

B) Importing our Meta-Data:

The most important step in any process is gathering all of our pre-requirements. In this stage we must import 2 of the most important requirements into our program, the data along with the required libraries for analysis.

i) Importing our Survey Data:

We will be implementing our project with the help of ' R '. In order for us to implement all our Data-sets which are stored in a single excel sheet into our program we must undergo the following procedure. The most important part is making sure that our sheet is properly saved in the folder of our wish. From here on out we will use the setwd() inbuilt within ' R ' so that we can set the path to the folder in which our sheet is saved in. For you to get a clear and easy understanding, let's suppose that we have saved our sheet on my desktop. In such a scenario we will use the following command.

setwd("C:/Users/mylaptop/Desktop")

Once we have used the setwd() function in ' R ' the path will be properly set to the designated folder. The next step involves selecting our data sheet so that our program can further access it for future analysis. To do this we will be using the inbuilt read.csv() function. This function however will return our file to have it stored in the form of a vector. To do this we must first declare a variable and then read our sheet into it. Since we are talking about the eating habits of people we've taken our variable as food. Let's take a look at how to read our sheet into our variable food…

food=read.csv("foodsurvey.csv", header = TRUE, stringsAsFactors = FALSE)

ii) Importing Required Libraries:

In order for us to further access the many tools within ' **R** ' we will have to first import the required libraries. One of the most important Libraries in ' **R** ' is Data Explorer. In order for us to do this we will run the following command…

library (DataExplorer)

This is one of the most important and widely used packages in **' R '** as it permits us to statistically analyse our data in many different methods.

It provides us many different functions that allow us to study the structure of our data as well as the variety of data collected. Further functions provided in this library permit us to plot the statistical analysis of our data in many different formats. The analysis of data is completely flexible with the desires of the user and can provide very accurate values and details.

C) Analysis & Presentation of our Data:

Now that we have imported all of the required data into our program we can now begin the process to begin its analysis. This step our process will not only contribute in studying the data but we will also look much deeper into how accurate as well as how volatile our data sets are.

i) Analysis of Our Survey:

One of the most important prerequisites of data analysis is to learn the accuracy as well as the completeness of our data-set. For us to do this we will use two important functions that are embedded in the *Data Explorer* package. The first function that we will be using is to learn about the structure of our Data-set. This displays all the questions asked in the form of nodes in a tree. All the questions that are represented as nodes are all connected to the root node which holds the integral structure of our tree. In order for us to identify the structure we will be using the following command in **' R '**...

plot_str(food)

On execution of this command we will get the following output Fig 1

The structure of data is:

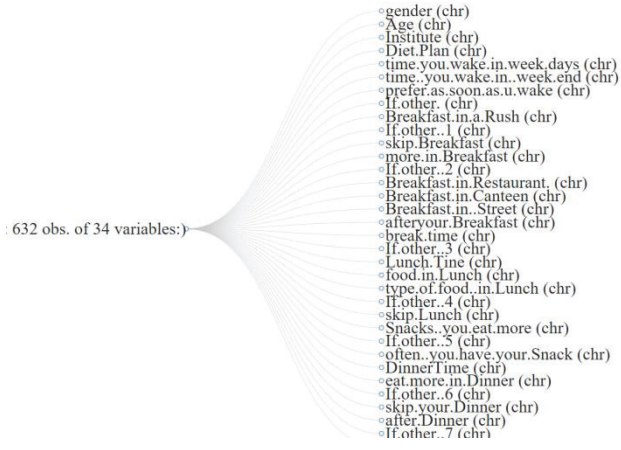


Fig. 1. Structure of data

Once we have established the structure of Data-Set we must also check to see how complete it is. In a survey people may sometimes chose to leave a few of the questions. In such a case our Data-set will not be a whole as there will be Null data.

Hence before we can begin our analysis we must first try to understand if our Data-set consists of missing data. In order to do this we will perform the following command in **‘ R ’**...

plot_missing(food)

Our Data-Set is 100% complete as from the output table Fig 2 we can observer that there is no missing data. If there were missing data it would of showed us the missing percentage of data. Suppose in a set of 10 entries only 9 were to answer the question, the amount of missing data would be 10%.

The plot of missing values of data is:

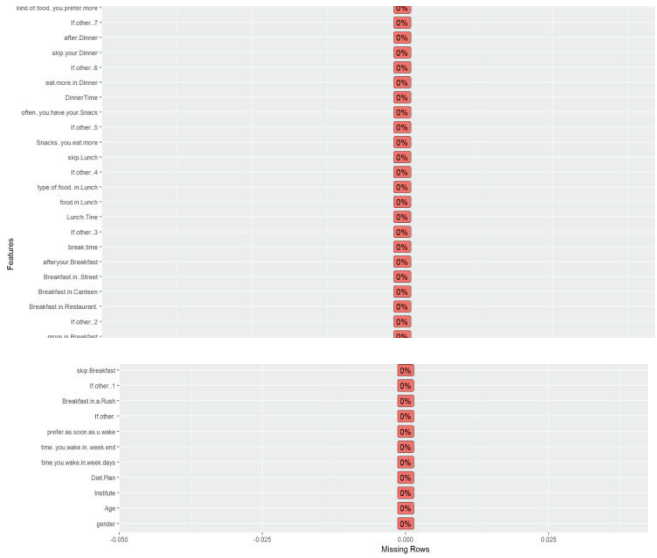


Fig. 2. Missing values of the data

b) Visual Representation of Data Analysis:

Now that we have verified our data it is time we begin our analysis. We can now plot our Dataset using the follwoing command…

plot_bar(food)

The output of this command will create a bar-graph for every question that we had our group of students answer. It will begin to analyse who chose what answers and will also provide you with accurate percentages for each option that is available to select. For you to get a small idea take a look at one of our outputs Fig 3

The visual representation of the data is:

Fig. 3. Visual representation of the given data

Using the plot function we can create multiple graphs of our choice. However instead of going through this whole process step by step we can use a single command which will help us complete all the precious steps in a much more fast and efficient manner. Report() is a function embedded within the Data-Explorer Package which completes all the previous steps that we have discussed about. It will make a complete analysis of our data which also includes the space/storage taken up by our Data-set and it also further counts the number of rows as well as columns. There are two more added functions to our report. It will also generate the ***Correlational Analysis*** along with the ***Principle Component Analysis***. We will further look into the features of these two further later on. The Report for our Data-Set can be generated using the following line…

create_report(food)

We will obtain the following Output Fig 4
The report that is created is:

Name	Value
Rows	632
Columns	34
Discrete columns	34
Continuous columns	0
All missing columns	0
Missing observations	0
Complete Rows	632
Total observations	21,488
Memory allocation	133.3 Kb

Fig. 4 (a) .Basic statics of Row Count

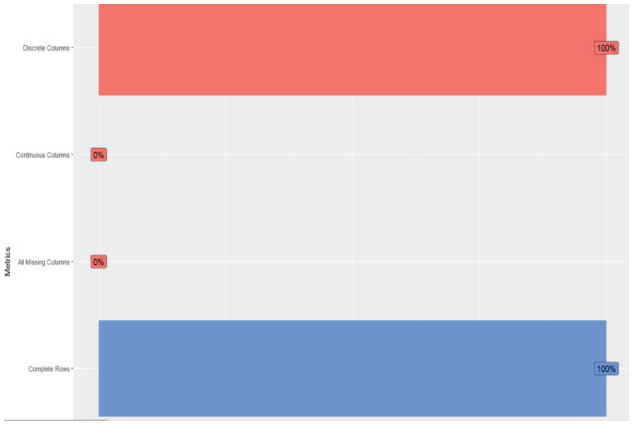


Fig. 4 (b).Basic statics of Percentages

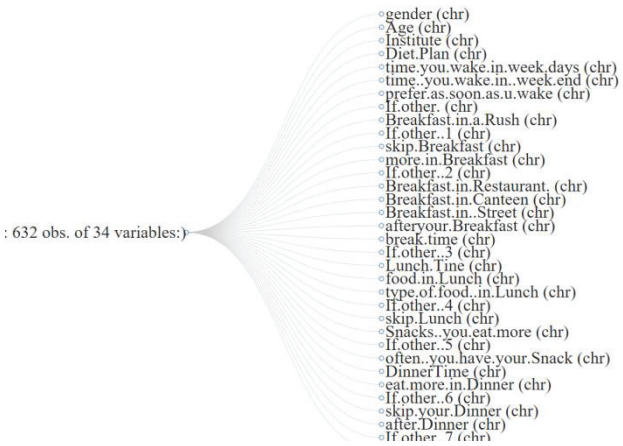


Fig. 4 (c).Structure of Data

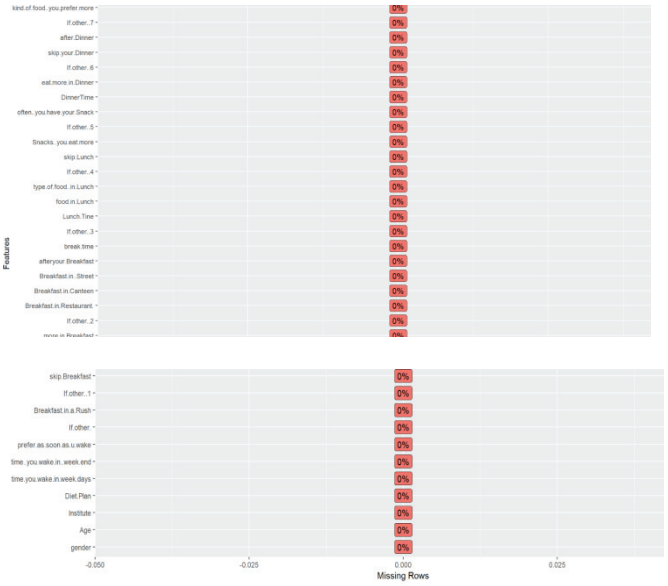
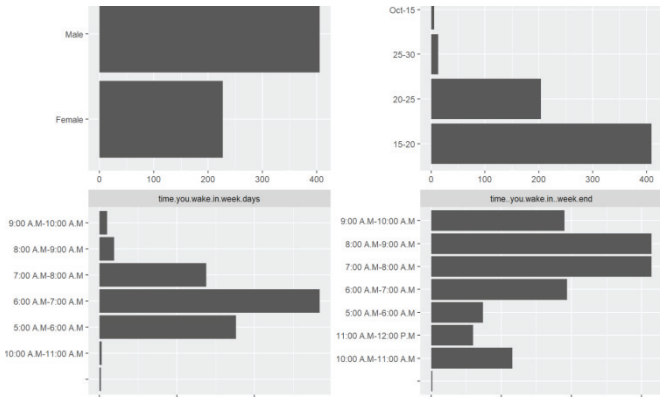
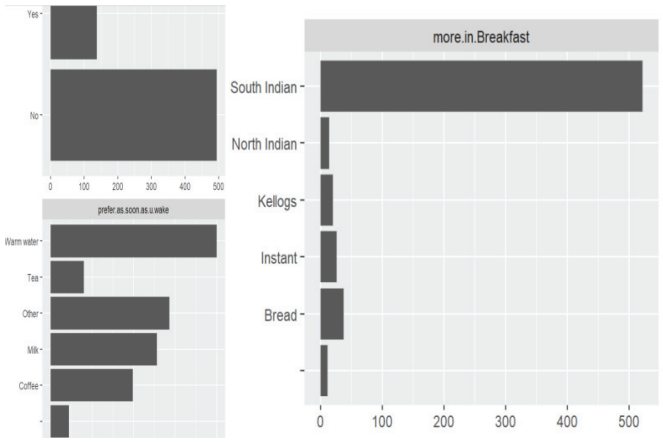
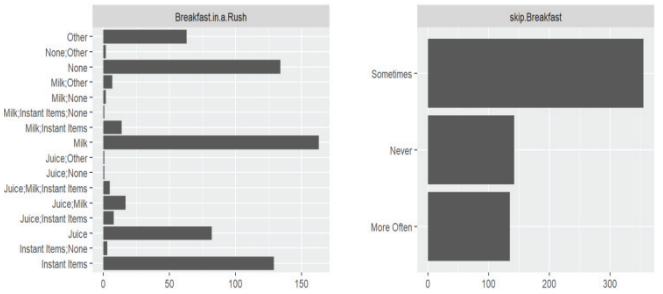
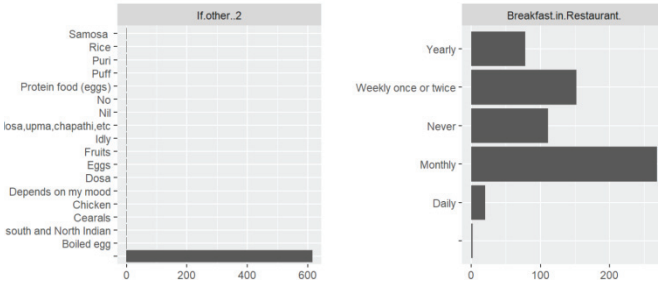


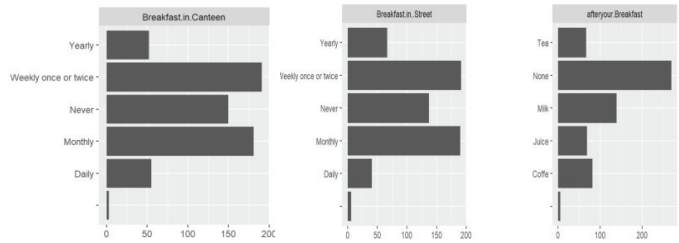
Fig. 4 (d).Missing values of Data

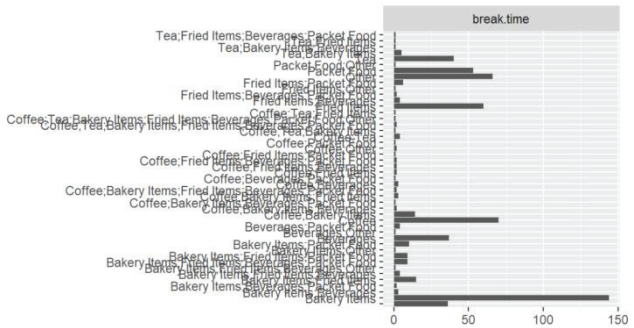


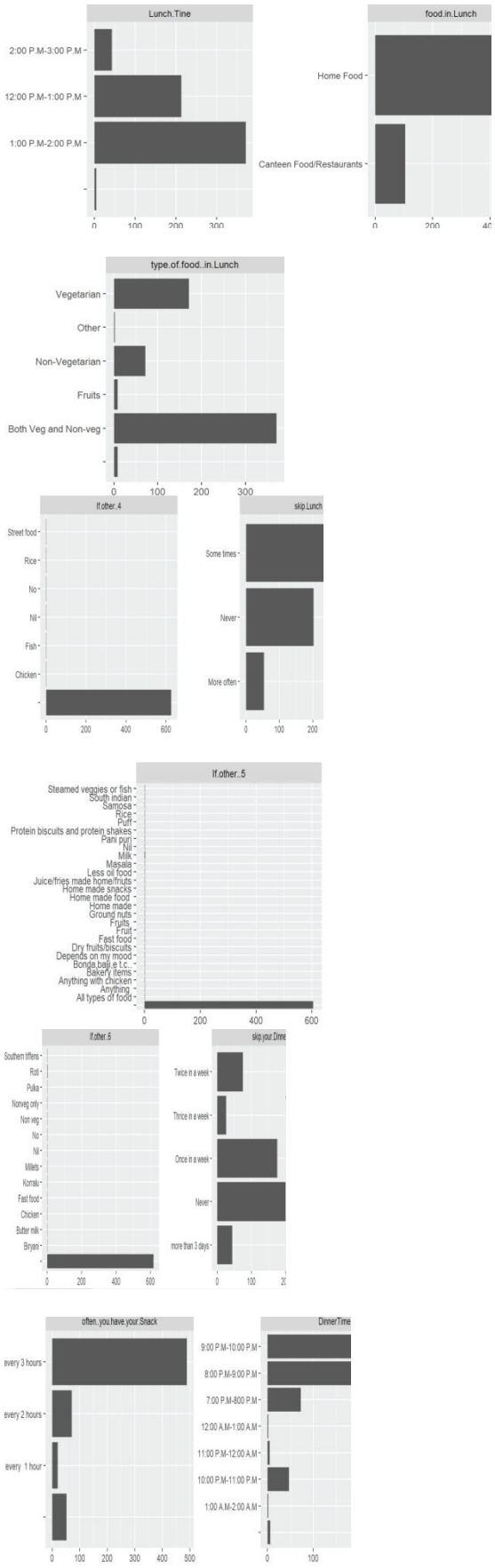












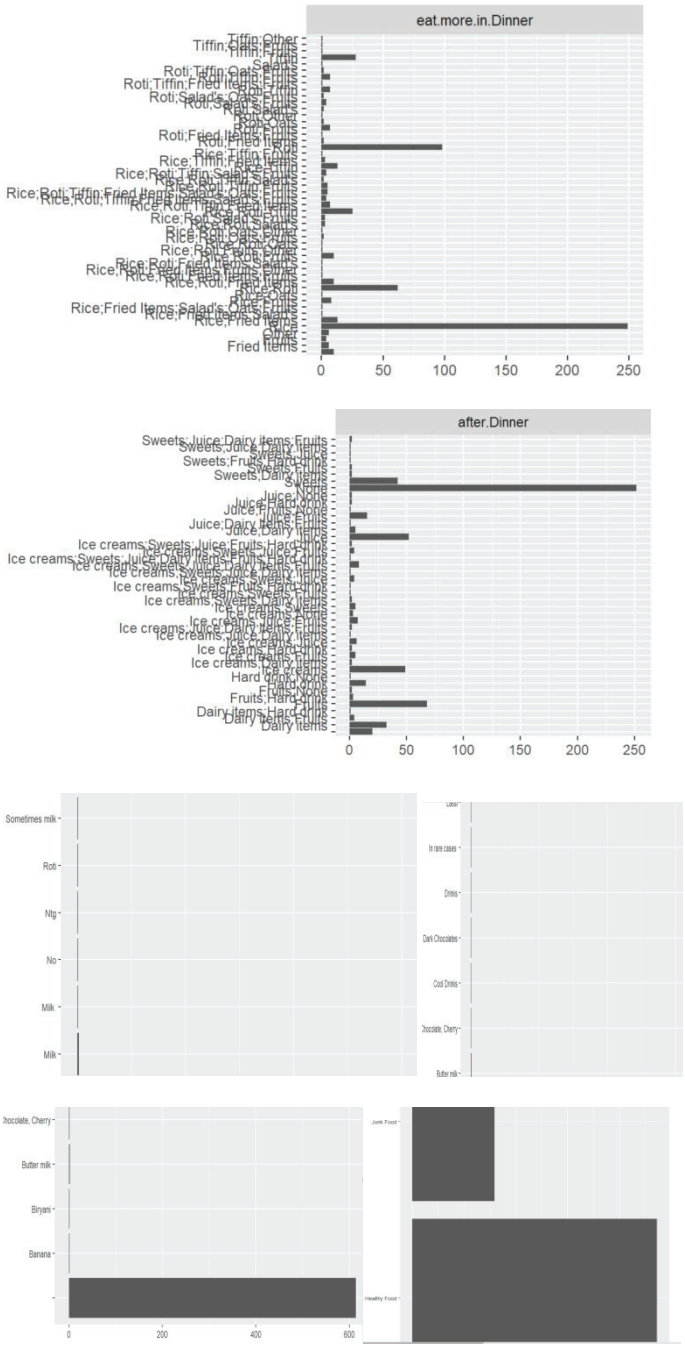


Fig. 4 (e).Analysis of Univariable Distribution of data

c) Correlational Analysis:

The name itself is quite self-explanatory. Correlational Analysis is basically the mapping of each variable that we have used. It will use a special correlation function to plot a graph which will help us understand how a single Variable in our Data-set is linked up or related with the other variables. This function lets us get a clear idea on all the variables and how dependent they may be. From this graph we can further make more analysed conclusions on our Data-Sets. Let's take a look

at how the Correlation function works both mathematically as well as graphically…

$$\mathrm{corr}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2\right]^{1/2}}$$

In the above formula n is the sample size of the data, $x_i$and y_i are the i^{th} data values and $\bar{x}$ and $\bar{y}$ are the mean values of the related attributes. The correlation function values always lies between -1 to +1.If the value is close to -1 it has strong negative correlation. If the value is close to 0 it has no correlation. If the values close to +1 it has strong positive correlation. In this relationship is identified based on the strength of dispersion of data points between two parameters and all the graphs that are drawn internally all are collected together and represented in a single graph in Fig. 5

The correlational graph is:

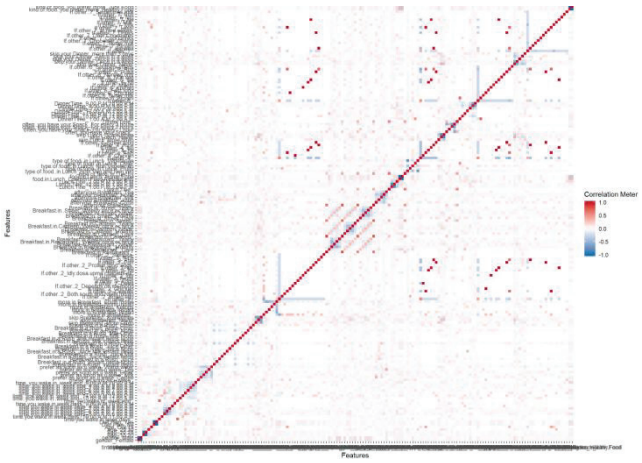


Fig. 5. The correlation Analysis of data

d). Principle Component Analysis:

The main plan of principal part analysis (PCA) is to scale back the spatial property of a knowledge set consisting of the many variables correlate with one another, either heavily or gently, while retaining the variation gift within the dataset, up to the utmost extent. The same is done by transforming the variables to a new set of variables, which are known as the principal parts (or merely,thePCs)and areaunit orthogonal,ordered such that the retention of variation gift within the original variables decreases as we have a tendency to move down within the order. So, during this approach, the first principal part retains most variation that was gift within the original parts. The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal.

The Principle component Analysis graph is:

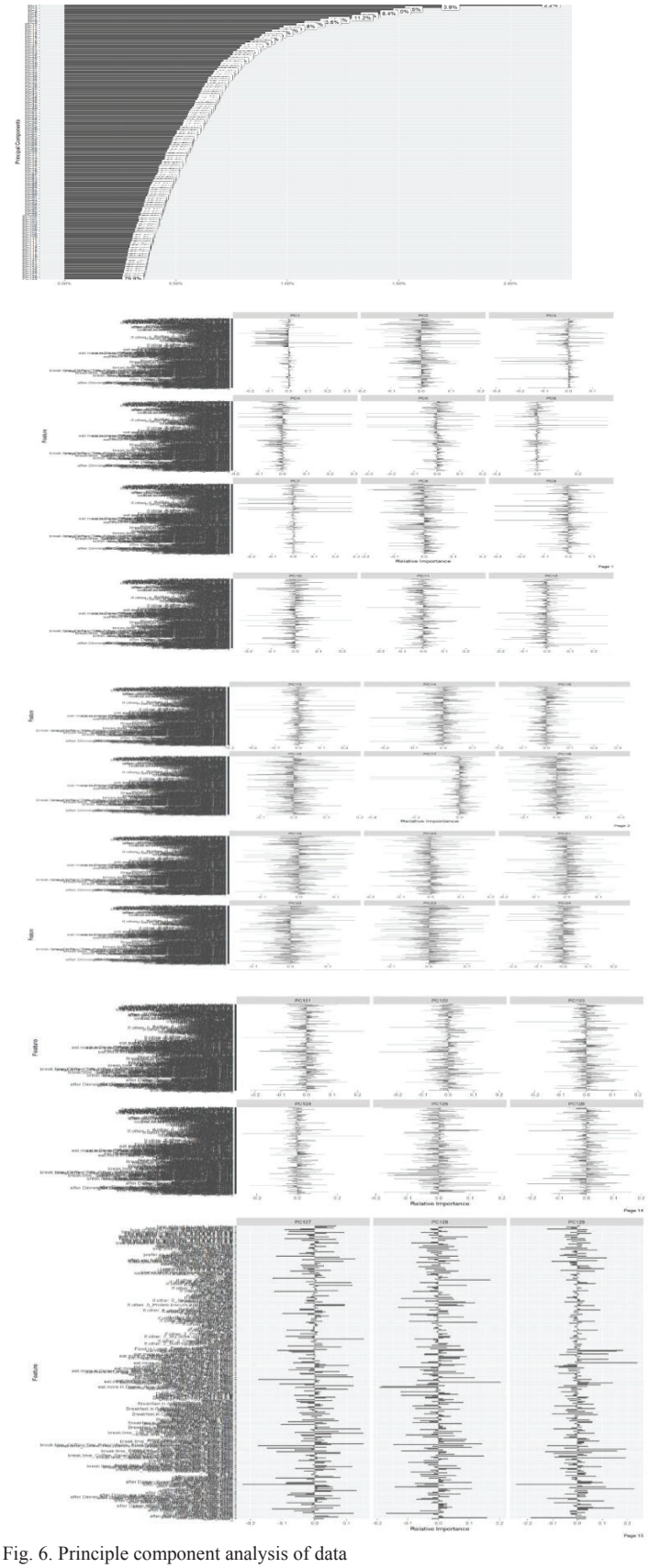


Fig. 6. Principle component analysis of data

III. CONCLUSION

We were able to successfully conclude that our modern generation lives in a state of dilemma. From our analysis we were able to learn that a huge majority of the people that participated in our survey prefer a healthy lifestyle. However most people according to their eating habits are actually on an unhealthy path yet quite unaware of it. Although they are leading an unhealthy lifestyle when asked how they overview their day to day habits they confidently answered that they lead a healthy lifestyle. We believe that it should be taken into our hands to further educate our modern generation on the importance of a healthy lifestyle along with what it truly means. They should be able to identify what's healthy and what's not. However no matter how we look at the analysis of our data we were able to conclude from our Data-set that eating healthy and maintaining a healthy lifestyle is preferred by most people…

REFERENCES

[1] Abraham S, Noriega Brooke R, Shin JY. College students eating habits and knowledge of nutritional requirements. J Nutr Hum Health. 2018;2(1):13-17

[2] J. Aravind, J. Dhalia Sweetlin. Nutrient Facts Analysis using Supervised Learning Approaches. 2017 Conference on Information and Communication Technology (CICT'17)

[3] NatnichaSuthumchai ; SirinThongsukh ; PacharamaiYusuksataporn ; Songsri Tangsripairoj; FoodForCare: An Android Application for Self-Care with Healthy Food2016 Fifth ICT International Student Project Conference (ICT-ISPC)

[4] Paul D Hatzigiannakoglou. Junk-Food Destroyer: Helping adolescents with Down syndrome to understand healthy eating through serious game; 2015 7th International Conference on Games and Virtual Worlds for Serious Applications (VS-Games).

[5] Moore PW,Burkhart KK,Jackson D.Drugs highly associated with infusionreactions reported using two different data-mining methodologies.J Blood Disorders Transf.2014;5:195.

[6] Kahraman, C., Cebeci, U. and D. Ruan, 2003, "Multi-attribute comparison of catering

[7] Liu, L., Bhattacharyya, S., Sclove, S.L., Chen, R. and W.J. Lattyak, 2001, "Data mining on

[8] Wiles NJ, Northstone K, Emmett P, Lewis G, Junk food diet and childhood behavioural problems: results from the ALSPAC cohort, 2009, 63(1), 491–498.

[9] Rang HP, Dale MM, Ritter JM, Moore PK. Pharmacology. 5th Ed. Delhi: Churchill Livingstone; 2006. P.394-400

[10] Medical devices: early warning of problems is hampered by sever underreporting. US General Accounting Office. GAO/PEMD 87-1; 1987.